PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 3/16, G10L 5/04	A1	(11) International Publication Number: WO 97/43707 (43) International Publication Date: 20 November 1997 (20.11.97)

(21) International Application Number: PCT/SE97/00584

(22) International Filing Date: 8 April 1997 (08.04.97)

(30) Priority Data:
9601812-2 13 May 1996 (13.05.96) SE

(71) Applicant: TELIA AB [SE/SE]; Mårbackagatan 11, S-123 86 Farsta (SE).

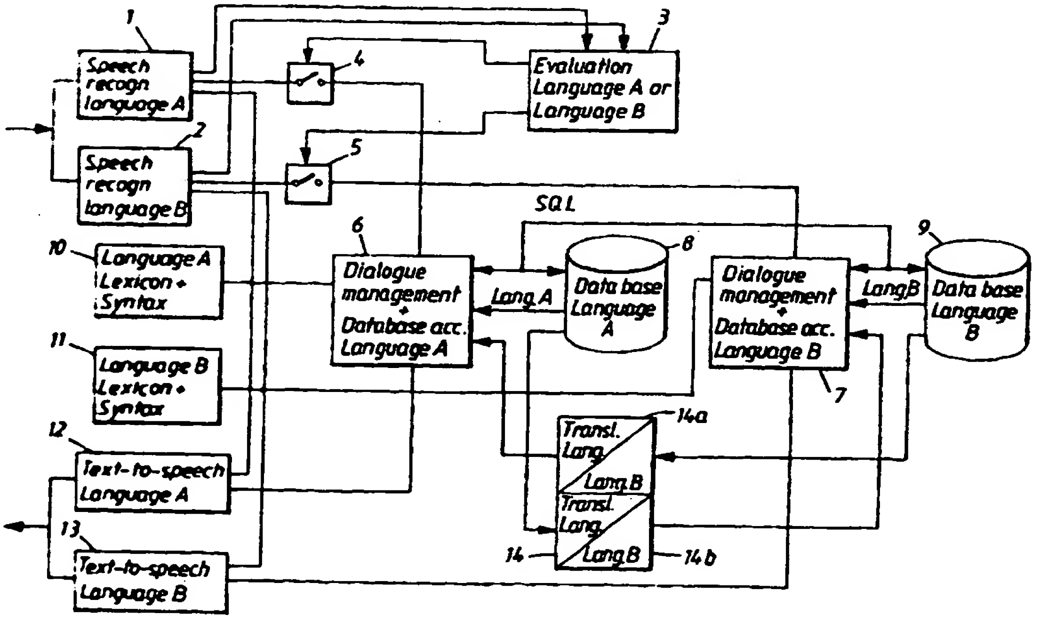
(72) Inventor: LYBERG, Bertil; Pingstvägen 5, S-610 72 Vagnhärad (SE).

(74) Agent: KARLSSON, Berne; Telia Research AB, Rudsjöterrassen 2, S-136 80 Haninge (SE).

(81) Designated States: NO, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: IMPROVEMENTS IN, OR RELATING TO, SPEECH-TO-SPEECH CONVERSION


1
Speech recogn language A
2
Speech recogn language B
4
5
3
Evaluation Language A or Language B
SQL
6
Dialogue management + Database acc. Language A
Lang A
8
Data base Language A
Dialogue management + Database acc. Language B
LangB
9
Data base Language B
7
10
Language A Lexicon + Syntax
11
Language B Lexicon + Syntax
12
Text-to-speech Language A
13
Text-to-speech Language B
Transl. Lang A Lang B
14a
Transl. Lang A Lang B
14
14b

(57) Abstract

A system and method for speech-to-speech conversion for providing spoken responses to speech inputs in at least two natural languages wherein speech inputs are recognised and interpreted in said at least two languages. The recognised speech inputs are evaluated to determine the language of the speech inputs, and a dialogue is undertaken with a database containing speech information data, in said at least two natural languages, to obtain data for the formulation of spoken responses to the speech inputs. The speech information data, obtained from the database, is then converted into spoken responses which exhibit the language characteristics of the respective speech inputs.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

IMPROVEMENTS IN, OR RELATING TO, SPEECH-TO-SPEECH CONVERSION

The present invention relates to a system and method for speech-to-speech conversion and to a voice responsive communication system including a speech-to-speech conversion system.

Known speech recognition systems which are adapted to provide spoken responses to speech inputs, include databases which contain speech information in many different languages and provide a recognition function for recognising and interpreting information in the languages concerned. However, the known speech recognition systems which may form part of speech-to-speech conversion systems, or the like, are dedicated to a single language, i.e. will only respond to speech inputs, e.g. spoken enquiries/questions, in the particular language which the system is adapted to handle and process.

In addition, the speech information data, which is stored in a database and used for the formulation of appropriate synthesised spoken responses to the speech inputs, is normally reproduced in a dialect which conforms to a standard national dialect. Thus, when there are significant differences between the dialect of the speech inputs and the standard national dialect, it may prove difficult, in certain circumstances, for the database of known speech-to-speech conversion systems to correctly interpret received speech information, i.e. the voice inputs to the system. It may also be difficult for the person making the voice inputs to fully understand the spoken response. Even if such responses are understandable to a recipient, it would be more user friendly if the dialect of the spoken response is the same as the dialect of the related voice input.

Also, with artificial reproduction of a spoken language, there is a need for the language to be reproduced naturally and with the correct accentuation. In particular, a word can have widely different meanings depending on language stress. Also, the meaning of one and the same sentence can depend on where the stress is placed. Furthermore, the stressing of sentences, or parts thereof, determines sections which are emphasised in the language and which may be of importance in determining the precise meaning of the spoken language.

The need for artificially produced speech to be as natural as possible and have the correct accentuation is of particular importance in voice responsive communication devices and/or systems which produce speech in various contexts. With known voice responsive arrangements, the reproduced speech is sometimes difficult to understand and interpret. There is, therefore, a need for a speech-to-speech conversion system in which the artificial speech outputs are natural, have the correct accentuation, and are readily understandable.

With languages having well developed sentence accent stress and/or pitch in individual words, identification of the natural meaning of the words/sentences is very difficult. The fact that stresses can be incorrectly placed increases the risk of misinterpretation, or that the meaning is completely lost for the listening party.

Thus, in order to overcome these difficulties, it would be necessary for a speech-to-speech conversion system to be capable of interpreting the received speech information, irrespective of language and/or dialect, and to match the language and/or dialect of speech outputs to that of the respective speech inputs. Also, in order to be able to determine the meaning of single words, or phrases, in an unambiguous manner in a spoken sequence, it would be necessary for the speech-to-speech conversion systems to be capable of determining, and taking account of, sentence accent and sentence stresses in the spoken sequence.

It is an object of the present invention to provide a system and method for speech-to-speech conversion which are adapted to recognize, interpret and process speech inputs in at least two natural languages and provide speech outputs, i.e. spoken responses, in the same language as the respective speech inputs.

It is another object of the present invention to provide a system and method for speech-to-speech conversion which are adapted to recognize, interpret and process speech inputs in at least two natural languages and provide speech outputs, i.e. spoken responses, in the same language and dialect as the respective speech inputs, the matching of the dialect being effected using prosody information and, more particularly, the fundamental tone curve of the speech inputs.

It is a further object of the present invention to provide a voice responsive communication system, including a speech-to-speech conversion system, operating in accordance with a speech-to-speech conversion method.

The invention provides, in a voice responsive communication system, a method for providing a spoken response to a speech input, said method including the steps of recognising and interpreting the speech input, and utilising the interpretation to obtain speech information data from a database for use in the formulation of the spoken response, characterised in that the database contains speech information data in at least two natural languages, in that said method is adapted to recognise and interpret speech inputs in said at least two languages and to provide spoken responses to speech inputs in said languages, and in that said method includes the further steps of evaluating a recognised speech input to determine the language of the input, effecting a dialogue with the database to obtain speech information data for the formulation of a spoken response in the language of the speech input, and converting the speech information data, obtained from the database, into said spoken response.

In a preferred method, separate databases may be used for each of said at least two languages, and dialogue may be effected with only that one of said databases which contains speech information data in the language of the input speech. However, in the event that at least part of the required speech information data for a spoken response is stored in another of said databases, the method may include the further steps of effecting a dialogue with said another database to obtain the required speech information data, translating the information data into the language of said one of the databases, combining the speech information data from the databases, and converting the combined speech information data into a spoken response in the language of the speech input.

The speech recognition and interpretation of a speech input may be effected in at least two natural languages. In this case, recognised parts, or sequences, of the speech input, resulting from the speech recognition and interpretation in the said at least two natural languages, are evaluated to determine the language of the speech input. The outcome of this evaluation process may be used to determine the database with which said dialogue is conducted to obtain the speech information data for a spoken response to the speech input.

The dialogue with a database, and/or between databases, may be effected using a database communication language, such as SQL (Structured Query Language).

In a preferred method, according to the present invention, the speech recognition and interpretation includes the steps of extracting prosody information, i.e. the fundamental tone curve, from a speech input, and obtaining dialect information from said prosody information, said dialect information being used in the conversion of said speech information data, obtained from said database, into a spoken response, the spoken responses being in the same language and dialect as the speech input. This preferred method, includes the further steps of determining the intonation pattern of the fundamental tone and thereby the maximum and minimum values of the fundamental tone curve and their respective positions; determining the intonation pattern of the fundamental tone curve of a speech model and thereby the maximum and minimum values of the fundamental tone curve and their respective positions; comparing the intonation pattern of the input speech with the intonation pattern of the speech model to identify a time difference between the occurrence of the maximum and minimum values of the fundamental tone curves of the incoming speech in relation to the maximum and minimum values of the fundamental tone curve of the speech model, the identified time difference being indicative of dialectal characteristics of the input speech. The time difference may be determined in relation to an intonation pattern reference point, for example, the point at which a consonant/vowel limit occurs.

The method, according to the present invention, may include the step of obtaining information on sentence accents from said prosody information.

The words in the speech model may be checked lexically, and the phrases in the speech model may be checked syntactically. The words and phrases which are not linguistically possible are excluded from the speech model. In addition, the orthography and phonetic transcription of the words in the speech model may be checked, the transcription information including lexically abstracted accent information, of type stressed syllables, and information relating to the location of secondary accent. The accent information may relate to tonal word accent I and accent II.

In addition, the method, according to the present invention, may use sentence accent information in the interpretation of the input speech.

The invention also provides a speech-to-speech conversion system for providing, at the output thereof, spoken responses to speech inputs in at least two natural languages, including speech recognition means for the speech inputs; interpretation means for interpreting the content of the recognised speech inputs, and a database containing speech information data for use in the formulation of said spoken responses, characterised in that the speech information data stored in the database is in the said at least two natural languages, in that the speech recognition and interpretation means are adapted to recognise and interpret speech inputs in said at least two natural languages, and in that the system further includes evaluation means for evaluating the recognised speech inputs and determining the language of the inputs, dialogue management means for effecting a dialogue with the database to obtain said speech information data in the language of a speech input, and text-to-speech conversion means for converting the speech information data, obtained from the database, into a spoken response.

The speech-to-speech conversion system, according to the present invention, which is adapted to receive speech inputs in two, or more, natural languages and to provide, at the output thereof, spoken responses in the language of the respective speech inputs, preferably includes, for each of the natural languages, speech recognition means, the inputs of each of the speech recognition means being connected to a common input for the system; speech evaluation means for determining, in dependence on the output of each of the speech recognition means, the language of a speech input; a database containing speech information data for use in the formulation of spoken responses in the language of the database; dialogue management means for connection to a respective speech recognition means, in dependence on the language of the speech input, said management means being adapted to interpret the content of recognised speech and, on the basis of the interpretation, to access, and obtain speech information data from, at least a respective one of the databases; and text-to-speech conversion means for converting the speech information data obtained by said management means into spoken responses to respective speech inputs.

The speech-to-speech conversion system may include separate databases for each of said at least two languages, and separate dialogue management means for each of the databases, each dialogue management means being adapted to effect a dialogue with, at least, a respective one of the databases. Also, each dialogue management means may be adapted to effect a dialogue with each of the databases. In this case, the system includes translation means for translating the speech information data output of each of the databases into the language of the other databases.

In the event that at least part of the required speech information data for a spoken response is stored in a database in a different language to that which is required for the spoken response, the speech information data may be obtained from said database and translated by said translation means into the required language for the spoken response. The translated speech information is then used either alone, or in combination with other speech information, by the dialogue management means to provide an output for application to the text-to-speech conversion means.

The speech-to-speech conversion system is preferably adapted to receive speech inputs in two languages, in which case, the system includes, for each of the two languages, a database, dialogue management means and translation means, in that each of the dialogue management means is adapted to communicate with each of the databases, the data output of each of the databases being connected directly to one of the dialogue management means and the other of the managemenτ means via a translation means.

The speech-to-speech conversion system preferably includes speech recognition and interpretation means for each of the said at least two natural languages, the inputs to the speech recognition and interpretation means being connected to a common system input. The recognised parts, or sequences, of the speech input, resulting from said speech recognition and interpretation in the said at least two natural languages, are evaluated by the evaluation means to determine the language of the speech input. The evaluation means may be used, in a preferred system, to select the database from which said speech information data will be obtained by said dialogue management means for the formulation of the spoken response to the speech input.

The speech recognition and interpretation means may include extraction means for extracting prosody information from the speech input, and means for obtaining dialectal information from said prosody information, said dialectal information being used by said text-to-speech conversion means in the conversion of said speech information data into the spoken response, the dialect of the spoken response being matched to that of the speech input. The prosody information extract from the speech input is the fundamental tone curve of the speech input.

The means for obtaining dialectal information from said prosody information may include first analysing means for determining the intonation pattern of the fundamental tone of the input speech and thereby the maximum and minimum values of the fundamental tone curve and their respective positions;

second analysing means for determining the intonation pattern of the fundamental tone curve of the speech model and thereby the maximum and minimum values of the fundamental tone curve and their respective positions; comparison means for comparing the intonation pattern of the input speech with the intonation pattern of the speech model to identify a time difference between the occurrence of the maximum and minimum values of the fundamental tone curves of the incoming speech in relation to the maximum and minimum values of the fundamental tone curve of the speech model, the identified time difference being indicative of dialectal characteristics of the input speech. The time difference may be determined in relation to an intonation pattern reference point, i.e. the point at which a consonant/vowel limit occurs.

The speech-to-speech conversion system may also include means for obtaining information on sentence accents from said prosody information.

The speech recognition means may include checking means for lexically checking the words in the speech model and for syntactically checking the phrases in the speech model, the words and phrases which are not linguistically possible being excluded from the speech model. The checking means may be adapted to check the orthography and phonetic transcription of the words in the speech model, in which case, the transcription information includes lexically abstracted accent information, of type stressed syllables, and information relating to the location of secondary accent. The accent information may relate to tonal word accent I and accent II.

The sentence accent information may be used in the interpretation of the content of the recognised input speech.

The sentence stresses may be determined and used in the interpretation of the content of the recognised input speech.

The invention further provides a voice responsive communication system which includes a speech-to-speech conversion system as outlined in the preceding paragraphs, or utilises a method as outlined in the preceding paragraphs for providing a spoken response to a speech input to the system.

The foregoing and other features of the present invention will be better understood from the following description, with reference to the single figure of the accompanying drawings, which illustrates, in the form of a block diagram, a speech-to-speech conversion system, according to the present invention.

The speech-to-speech conversion system, according to the present invention, is adapted to provide, at the output thereof, spoken responses to speech inputs in at least two natural languages. The language characteristics of the spoken responses, for example, dialect, sentence accent and sentence stresses, are matched, by the present invention, to those of the input speech to provide natural speech outputs which can be readily understood, have the correct accentuation and give rise to a user friendly system. It will be seen from the following description that the matching of the language characteristics is achieved by extracting prosody information from the speech input, i.e. the fundamental tone curve of the speech input, and using the prosody information to determine dialectal, sentence accent and sentence stressing, information for use in the formulation of the spoken responses.

The speech-to-speech conversion system may, therefore, be used in many applications, for example, in voice responsive communication systems to effect a dialogue between a user of the system and a database which forms part of the system's speech recognition unit and which contains speech information data for the formulation of spoken responses to spoken questions/enquiries from users of the system. Such voice responsive communication systems could be used in telecommunications, or banking, or security, etc., to provide a readily understandable, user friendly, system.

The speech-to-speech conversion system, illustrated in the single figure of the accompanying drawings, is adapted to provide, at the output thereof, spoken responses to speech inputs in two natural languages, i.e. languages A and B, which may be any natural language, for example Swedish and English.

As shown in the accompanying drawing, the system includes speech recognition and interpretation units 1 and 2, respectively, for the languages A and B. The inputs of the units 1 and 2 are connected to a common input for the system. The speech recognition and interpretation units 1 and 2 are used to recognise and interpret the content of the speech input in a manner to be subsequently outlined.

An output of each of the units 1 and 2 are connected to separate inputs of an evaluation unit 3 which is adapted to evaluate the recognised speech inputs and determine the language of the inputs, i.e. language A, or language B.

The system of the present invention also includes two switching units 4 and 5, the inputs of which are respectively connected to an output of the speech recognition and interpretation units 1 and 2. Operation of the switching units 4 and 5 is controlled, in a manner to be subsequently outlined, by the evaluation unit 3, i.e. the control inputs to the units 4 and 5 are respectively connected to separate outputs of the evaluation unit 3.

The outputs of the switching units 4 and 5 are respectively connected to an input of dialogue management units 6 and 7. It will be seen from subsequent description that the dialogue management units 6 and 7 are used to effect a dialogue with database units 8 and 9 to obtain speech information data, in the language of a speech input, for use in the formulation of the spoken responses.

A lexicon and syntax unit 10 for the language A is connected to another output of the speech recognition and interpretation unit 1, to the dialogue management unit 6 and to an input of a text-to-speech conversion unit 12.

A lexicon and syntax unit 11 for the language B is connected to another output of the speech recognition and interpretation unit 2, to the dialogue management unit 7 and to an input of a text-to-speech conversion unit 13.

The text-to-speech conversion units 12 and 13 are also respectively connected, at another input thereof, to an output of dialogue management units 6 and 7.

The outputs of the text-to-speech conversion units 12 and 13 are connected to a common speech output for the system.

As shown in the accompanying drawing, there is a two way communication path between the dialogue management unit 6 and database unit 8, and between the dialogue management unit 7 and database unit 9. These communication paths are used to effect, in a manner to be subsequently outlined, a dialogue between the respective management and database units to obtain speech information data for use in the formulation of the spoken responses. The two way communication paths are interconnected to enable a dialogue to be undertaken between management unit 6 and database unit 9 and/or between management unit 7 and database unit 8. In practice, the dialogue with a database unit, and/or between databases units, is effected using a database communication language, such as SQL (Structured Query Language).

A translation unit 14 is provided for translating language A into language B and vice versa. It will be seen from the accompanying drawing that one section 14a of the translation unit 14 has an input for language B which is connected to an output of database unit 9, and an output for language A which is connected to an input of dialogue management unit 6. Another section 14b of the translation unit 14 has an input for language A which is connected to an output of database unit 8, and an output for language B which is connected to an input of dialogue management unit 7.

The manner in which the speech-to-speech conversion system is adapted to receive speech inputs in natural languages A and B, and to provide, at the output thereof, spoken responses in the language of the respective speech input, is outlined in the following paragraphs.

A speech input to the speech-to-speech conversion system which can be in either language A, or language B, is recognised and interpreted by each of the speech recognition and interpretation units 1 and 2, in association with the respective lexicon and syntax units 10 and 11, i.e. using statistically-based speech recognition and language modelling techniques, and ensuring that the recognised words and/or word combinations which are used to form a model of the speech input, are acceptable both lexically and syntactically. The purpose of the lexicon/syntax checks is to identify and exclude any words from the speech model which do not exist in the language concerned, and/or any phrase whose syntax does not correspond with the language concerned.

The speech models respectively created by the units 1 and 10, and the units 2 and 11, are applied to, and evaluated by, the evaluation unit 3 which determines which of the languages A and B is most probable for the speech input. This evaluation is effected on the basis of probability, i.e. the probability that the speech input is one, or other, of the languages A and B, the differences between the speech models, and whether the language modelling for one, or other, of the languages has been successfully completed. The greater the difference between the language characteristics of languages A and B, the easier will be the task of the evaluation unit 3.

Depending on the outcome of the evaluation by the unit 3, i.e. the selected language of the speech input, one of the switching units 4 and 5 will be activated to connect the speech recognition and interpretation unit for the selected language to the corresponding dialogue management unit.

If it is assumed, for the purpose of this description, that language A has been selected as the most probable language for the speech input, then switching unit 4 will be activated and the output of speech recognition and interpretation unit 1 will be connected to an input of dialogue management unit 6. Thus, the switching unit 5 will remain in a deactivated state and no connection will, therefore, be made between the dialogue management unit 9 and the speech recognition and interpretation unit 2.

In the next stage of the speech-to-speech conversion process, the management unit 6 enters into a linguistic dialogue with the database unit 8, on the basis of the speech model of the speech input, to obtain speech information data for the formulation of a spoken response to the speech input. The speech information data, selected as result of this dialogue, is transferred via the management unit 6 to an input of the text-to-speech converting unit 5 for the formulation of a spoken response. It will be seen from subsequent description that the language characteristics of the spoken response is matched, as far as possible, to the language characteristics of the speech input.

In the event that at least part of the required speech information data for a spoken response is not stored in the database unit 6, but may be stored in the database unit 9, the dialogue management unit 6 enters into a dialogue with the database unit 9 to obtain the required speech information data. If the required speech information data is stored in database unit 9, it is accessed and transferred to the dialogue management unit 6 via section 14a of the translation unit 14, i.e. is translated from language B into language A. The translated speech information data is then used either alone, or in combination with speech information data obtained from the database unit 8, to formulate the spoken response, i.e. converted by the text-to-speech conversion unit 12 into the spoken response.

Clearly, if language B, rather than language A, is selected by the evaluation unit 3 as the language of the speech input, then the units 7, 9 and 13 would be used, in the same manner as outlined above for the units 6, 8 and 12, for the formulation of the spoken response. Any information that may be required from the database unit 8 would be accessed by, and transferred to, the dialogue management unit 7, the translation of the transferred information data being effected by section 14b of the translation unit 14.

The recognition and interpretation of speech can give rise to technical problems and if these problems are not overcome, then difficulties will be experienced in obtaining a correct and meaningful interpretation of the speech inputs. In particular, if the recognition and interpretation of the speech inputs is incorrect, then it will be extremely difficult for the evaluation unit 3 to determine the language of the speech inputs and it will not, therefore, be possible to provide proper spoken responses to the speech inputs.

Thus, in accordance with the present invention, these problems are overcome by extracting prosody information from the speech inputs and using this information to determine, in a manner to be subsequently outlined, dialectal, sentence accent, and sentence stressing, information for use in the recognition and interpretation process and in the formulation of the spoken responses.

The extraction of the prosody information, i.e. the fundamental tone curve, from the speech input is effected by prosody extraction means (not illustrated) which form part of the speech recognition and interpretation units 1 and 2. These units also include means (not illustrated) for obtaining dialectal information from the prosody information.

Thus, with the present invention, the speech recognition and interpretation units 1 and 2 are adapted to operate, in a manner well known to persons skilled in the art, to recognise and interpret the speech inputs to the system. The speech recognition and interpretation units 1 and 2 may, for example, operate by using a Hidden Markov model, or an equivalent speech model. In essence, the function of the units 1 and 2 is to convert speech inputs to the system into a form, which is a faithful representation of the content of the speech inputs, and which is suitable for evaluation by the evaluation unit 3 and use by the dialogue management units 6 and 7. In other words, the content of the textual information data, at the output of each of the speech recognition and interpretation units 1 and 2, must be:

- an accurate representation of the speech input; and
- be usable by the database management units 6 and 7 to respectively access, and extract speech information data from, the database units 8 and 9, for use in the formulation of a synthesised spoken response, i.e. by a respective one of the text-to-speech conversion units 12 and 13.

In practice, the recognition and interpretation process would, in essence, be effected by identifying a number of phonemes from a segment of the speech input which are combined into allophone strings, the phonemes being interpreted as possible words, or word combinations, to establish a model of the speech. The established speech model will have word and sentence accents according to a standardised pattern for the language of the input speech.

The information, concerning the recognised words and word combinations, generated by the speech recognition and interpretation unit 1 and 2, is checked, in a manner as outlined above, both lexically and syntactically. In practice, this would be effected using a lexicon with orthography and transcription.

Thus, in accordance with the present invention, the speech recognition and interpretation units 1 and 2 ensure that only those words, and word combinations, which are found to be acceptable both lexically and syntactically, are used to create a model of the input speech. In practice, the intonation pattern of the speech model is a standardised intonation pattern for the language concerned, or an intonation pattern which has been established by training, or explicit knowledge, using a number of dialects of the language concerned.

As stated above, the prosody information, i.e. the fundamental tone curve, extracted from the input speech by the extraction unit 3, can be used to obtain dialectal, sentence accent and sentence stressing, information, for use by the speech-to-speech conversion system and method of the present invention. In particular, the dialectal information can be used by the speech-to-speech conversion system and method to match the dialect of the output speech to that of the input speech and the sentence accent and stressing information can be used in the recognition and interpretation of the input speech.

In accordance with the present invention, the means for obtaining dialectal information from the prosody information includes;

- first analysing means for determining the intonation pattern of the fundamental tone of the input speech and thereby the maximum and minimum values of the fundamental tone curve and their respective positions;

- second analysing means for determining the intonation pattern of the fundamental tone curve of the speech model and thereby the maximum and minimum values of the fundamental tone curve and their respective positions; and

- comparison means for comparing the intonation pattern of the input speech with the intonation pattern of the speech model to identify a time difference between the occurrence of the maximum and minimum values of the fundamental tone curves of the incoming speech in relation to the maximum and minimum values of the fundamental tone curve of the speech model, the identified time difference being indicative of the dialectal characteristics of the input speech.

The time difference, referred to above, may be determined in relation to an intonation pattern reference point.

In the Swedish language, the difference, in terms of intonation pattern, between different dialects can be described by different points in time for word and sentence accent, i.e. the time difference can be determined in relation to an intonation pattern reference point, for example, the point at which a consonant/vowel limit occurs.

Thus, in a preferred arrangement for the present invention, the reference against which the time difference is measured, is the point at which the consonant/vowel boundary, i.e. the CV-boundary, occurs.

The identified time difference which, as stated above, is indicative of the dialect in the input speech, i.e. the spoken language, is applied to the text-to-speech conversion units 12 and 13 to enable the intonation pattern, and thereby the dialect, of the speech output of the system to be corrected so that it corresponds to the intonation pattern of the corresponding words and/or phrase of the input speech. Thus, this corrective process enables the dialectal information in the input speech to be incorporated into the output speech.

As stated above, the fundamental tone curve of the speech model is based on information resulting from the lexical (orthography and transcription) and syntactic checks. In addition, the transcription information includes lexically abstracted accent information, of type stressed syllables, i.e. tonal word accents I and II, and information relating to the location of secondary accents, i.e. information given, for instance, in dictionaries. This information can be used to adjust the recognition pattern of the speech recognition model, for example, the Hidden Markov model, to take account of the transcription information. A more exact model of the input speech is, therefore, obtained during the interpretation process.

A further consequence of this speech model corrective process is that, through time, the speech model will have an intonation pattern which has been established by a training process.

Also, with the system and method of the present invention, the speech model is compared with a spoken input sequence, and any difference there between can be determined and used to bring the speech model into conformity with the spoken sequence and/or to determine stresses in the spoken sequence.

In addition, the identification of the stresses in a spoken sequence makes it possible to determine the precise meaning of the spoken sequence in an unambiguous manner. In particular, relative sentence stresses can be determined by classifying the ratio between variations and declination of the fundamental tone curve, whereby emphasised sections, or individual words can be determined. In addition, the pitch of the speech can be determined from the declination of the fundamental tone curve.

Thus, in order to take account of sentence stresses in the recognition and interpretation of the speech inputs to the speech-to-speech conversion system of the present invention, the prosody extraction means and the associated speech recognition and interpretation unit, for each of the languages A and B, is adapted to determine:

- a first ratio between the variation and declination of the fundamental tone curve of the input speech;
- a second ratio between the variation and declination of the fundamental tone curve of the speech model; and
- comparing the first and second ratios, any identified difference being used to determine sentence accent placements.

Furthermore, classification of the ratio between the variation and declination of the fundamental tone curve, makes it possible to identify/determine relative sentence stresses, and emphasised sections, or words.

Also, the relation between the variation and declination of the fundamental tone curve can be utilised to determine the dynamic range of the fundamental tone curve.

The information obtained in respect of the fundamental tone curve concerning dialect, sentence accent and stressing, can be used for the interpretation of speech inputs by the units 1 and 2, i.e. the information can be used, in the manner outlined above, to obtain a better understanding of the content of the input speech and bring the intonation pattern of the speech model into conformity with the input speech.

Since the corrected speech model exhibits the language characteristics (including dialect information, sentence accent and stressing) of the input speech, it can be used to give an increased understanding of the input speech and increase the probability that the evaluation unit 3 will select the correct language of the speech inputs. The corrected speech models can also be used by the database management units 6 and 7 to obtain the required speech information data from the database units 8 and 9 for the formulation of a response to a voice input to the speech-to-speech conversion system.

The ability to readily interpret different dialects in a language using fundamental tone curve information, is of some significance because such interpretations can be effected without having to train the speech recognition system. The result of this is that the size, and thereby cost, of a speech recognition system, made in accordance with the present invention, can be much smaller than would be possible with known systems. These are, therefore, distinct advantages over known speech recognition systems.

The system is, therefore, adapted to recognise and accurately interpret the content of speech inputs in two, or more, natural languages and to match the language characteristics, e.g. dialect, of the voice responses to those of the voice inputs. This process provides a user friendly system because the language of the man-machine dialogue is in accordance with the dialect of the user concerned.

The present invention is not limited to the embodiments outlined above, but can be modified within the scope of the appended patent claims and the inventive concept.

CLAIMS

1. In a voice responsive communication system, a method for providing a spoken response to a speech input, said method including the steps of recognising and interpreting the speech input, and utilising the interpretation to obtain speech information data from a database for use in the formulation of the spoken response, characterised in that the database contains speech information data in at least two natural languages, in that said method is adapted to recognise and interpret speech inputs in said at least two languages and to provide spoken responses to speech inputs in said languages, and in that said method includes the further steps of evaluating a recognised speech input to determine the language of the input, effecting a dialogue with the database to obtain speech information data for the formulation of a spoken response in the language of the speech input, and converting the speech information data, obtained from the database, into said spoken response.

2. A method as claimed in claim 1, characterised in that separate databases are used for each of said at least two languages.

3. A method as claimed in claim 2, characterised in that said dialogue is effected with only that one of said databases which contains speech information data in the language of the input speech.

4. A method as claimed in claim 2, characterised in that said dialogue is effected with that one of said databases which contains speech information in the language of the input speech, and in that, in the event that at least part of the required speech information data for a spoken response is stored in another of said databases, said method includes the further steps of effecting a dialogue with said another of the

databases to obtain the required speech information data, translating the information data into the language of said one of the databases, combining the speech information data from the databases, and converting the combined speech information data into a spoken response in the language of the speech input.

5. A method as claimed in any one of the preceding claims, characterised in that speech recognition and interpretation of a speech input is effected in at least two natural languages.

6. A method as claimed in claim 5, characterised in that recognised parts, or sequences, of the speech input, resulting from said speech recognition and interpretation in the said at least two natural languages, are evaluated to determine the language of the speech input.

7. A method as claimed in claim 6, characterised in that the outcome of the evaluation process is used to determine the database with which the said dialogue is conducted to obtain the speech information data for a spoken response to the speech input.

8. A method as claimed in any one of the preceding claims, characterised in that the dialogue with a database and/or between databases is effected using a database communication language, such as SQL (Structured Query Language).

9. A method as claimed in any one of the preceding claims characterised in that said speech recognition and interpretation includes the steps of extracting prosody information from a speech input, and obtaining dialect information from said prosody information, said dialect information being used in the conversion of said speech information data, obtained from said database, into a spoken response, the spoken responses being in the same language and

dialect as the speech input.

10. A method as claimed in claim 9, characterised in that the prosody information extract from the speech input is the fundamental tone curve of the speech input.

11. A method as claimed in claim 10, characterised by the steps of determining the intonation pattern of the fundamental tone curve of the input speech and thereby the maximum and minimum values of the fundamental tone curve and their respective positions; determining the intonation pattern of the fundamental tone curve of a speech model and thereby the maximum and minimum values of the fundamental tone curve and their respective positions; comparing the intonation pattern of the input speech with the intonation pattern of the speech model to identify a time difference between the occurrence of the maximum and minimum values of the fundamental tone curves of the incoming speech in relation to the maximum and minimum values of the fundamental tone curve of the speech model, the identified time difference being indicative of dialectal characteristics of the input speech.

12. A method as claimed in claim 11, characterised in that the time difference is determined in relation to an intonation pattern reference point.

13. A method as claimed in claim 12, characterised in that the intonation pattern reference point, against which the time difference is measured, is the point at which a consonant/vowel limit occurs.

14. A method as claimed in any one of the claims 9 to 13, characterised by the step of obtaining information on sentence accents from said prosody information.

15. A method as claimed in claim 14, characterised in that the

words in the speech model are checked lexically, in that the phrases in the speech model are checked syntactically, in that the words and phrases which are not linguistically possible are excluded from the speech model, in that the orthography and phonetic transcription of the words in the speech model are checked, and in that the transcription information includes lexically abstracted accent information, of type stressed syllables, and information relating to the location of secondary accents.

16. A method as claimed in claim 15, characterised in that the accent information relates to tonal word accent I and accent II.

17. A method as claimed in any one of claims 14 to 16, characterised by the step of using said sentence accent information in the interpretation of the input speech.

18. A voice responsive communication system which utilises a method as claimed in any one of the preceding claims for providing a spoken response to a speech input to the system.

19. A speech-to-speech conversion system for providing, at the output thereof, spoken responses to speech inputs in at least two natural languages, including speech recognition means for the speech inputs; interpretation means for interpreting the content of the recognised speech inputs, and a database containing speech information data for use in the formulation of said spoken responses, characterised in that the speech information data stored in the database is in the said at least two natural languages, in that the speech recognition and interpretation means are adapted to recognise and interpret speech inputs in said at least two natural languages, and in that the system further includes evaluation means for evaluating the recognised speech inputs and determining the language of the inputs, dialogue management means for effecting

a dialogue with the database to obtain said speech information data in the language of a speech input, and text-to-speech conversion means for converting the speech information data, obtained from the database, into a spoken response.

20. A speech-to-speech conversion system as claimed in claim 19, characterised in that the system is adapted to receive speech inputs in two, or more, natural languages and to provide, at the output thereof, spoken responses in the language of the respective speech inputs, and in that the system includes, for each of the natural languages, speech recognition means, the inputs of each of the speech recognition means being connected to a common input for the system; speech evaluation means for determining, in dependence on the output of each of the speech recognition means, the language of a speech input; a database containing speech information data for use in the formulation of spoken responses in the language of the database; dialogue management means for connection to a respective speech recognition means, in dependence on the language of the speech input, said management means being adapted to interpret the content of recognised speech and, on the basis of the interpretation, to access, and obtain speech information data from, at least a respective one of the databases; and text-to-speech conversion means for converting the speech information data obtained by said management means into spoken responses to respective speech inputs.

21. A speech-to-speech conversion system as claimed in claim 19, characterised in that the system includes separate databases for each of said at least two languages.

22. A speech-to-speech conversion system as claimed in claim 21, characterised in that the system includes separate dialogue management means for each of the databases, each dialogue management means being adapted to effect a dialogue with, at least, a respective one of the databases.

23 A speech-to-speech conversion system as claimed in claim 22, characterised in that each dialogue management means is adapted to effect a dialogue with each of the databases.

24. A speech-to-speech conversion system as claimed in claim 23, characterised in that the system includes translation means for translating the speech information data output of each of the databases into the language(s) of the other databases.

25. A speech-to-speech conversion system as claimed in claim 24, characterised in that, in the event that at least part of the required speech information data for a spoken response is stored in a database in a different language to that which is required for the spoken response, said information is obtained from said database and translated by said translation means into the required language for the spoken response, and in that the translated speech information is used either alone, or in combination, with other speech information by the dialogue management means to provide an output for application to the text-to-speech conversion means.

26. A speech-to-speech conversion system as claimed in claim 25, characterised in that the system is adapted to receive speech inputs in two languages, in that the system includes, for each of the two languages, a database, dialogue management means and translation means, in that each of the dialogue management means is adapted to communicate with each of the databases, and in that the data output of each of the databases is connected directly to one of the dialogue management means and to the other of the management means via a translation means.

27. A speech-to-speech conversion system as claimed in any one of the claims 19 to 26, characterised in that the system includes speech recognition and interpretation means for each of the said at least two natural languages, the inputs to the

speech recognition and interpretation means being connected to a common system input.

28. A speech-to-speech conversion system as claimed in claim 27, characterised in that recognised parts, or sequences, of the speech input, resulting from said speech recognition and interpretation in the said at least two natural languages, are evaluated by the evaluation means to determine the language of the speech input.

29. A speech-to-speech conversion system as claimed in claim 28, characterised in that the output of the evaluation means is used to select the database from which said speech information data will be obtained by said dialogue management means for the formulation of the spoken response to the speech input.

30. A speech-to-speech conversion system as claimed in any one of claims 19 to 29, characterised in that the dialogue with a database, and/or between databases, is effected using a database communication language, such as SQL (Structured Query Language).

31. A speech-to-speech conversion system as claimed in any one of the preceding claims characterised in that said speech recognition and interpretation means include extraction means for extracting prosody information from the speech input, and means for obtaining dialectal information from said prosody information, said dialectal information being used by said text-to-speech conversion means in the conversion of said speech information data into the spoken response, the dialect of the spoken response being matched to that of the speech input.

32. A speech-to-speech conversion system as claimed in claim 31, characterised in that the prosody information extract from

the speech input is the fundamental tone curve of the speech input.

33. A speech-to-speech conversion system as claimed in claim 32, characterised the means for obtaining dialectal information from said prosody information includes first analysing means for determining the intonation pattern of the fundamental tone of the input speech and thereby the maximum and minimum values of the fundamental tone curve and their respective positions; second analysing means for determining the intonation pattern of the fundamental tone curve of the speech model and thereby the maximum and minimum values of the fundamental tone curve and their respective positions; comparison means for comparing the intonation pattern of the input speech with the intonation pattern of the speech model to identify a time difference between the occurrence of the maximum and minimum values of the fundamental tone curves of the incoming speech in relation to the maximum and minimum values of the fundamental tone curve of the speech model, the identified time difference being indicative of dialectal characteristics of the input speech.

34. A speech-to-speech conversion system as claimed in claim 33, characterised in that the time difference is determined in relation to an intonation pattern reference point.

35. A speech-to-speech conversion system as claimed in claim 34, characterised in that the intonation pattern reference point, against which the time difference is measured, is the point at which a consonant/vowel limit occurs.

36. A speech-to-speech conversion system as claimed in any one of the claims 31 to 35, characterised in that the system further includes means for obtaining information on sentence accents from said prosody information.

37. A speech-to-speech conversion system as claimed in claim

36, characterised in that the speech recognition means includes checking means for lexically checking the words in the speech model and for syntactically checking the phrases in the speech model, the words and phrases which are not linguistically possible being excluded from the speech model, in that the checking means are adapted to check the orthography and phonetic transcription of the words in the speech model, in that the transcription information includes lexically abstracted accent information, of type stressed syllables, and information relating to the location of secondary accent.

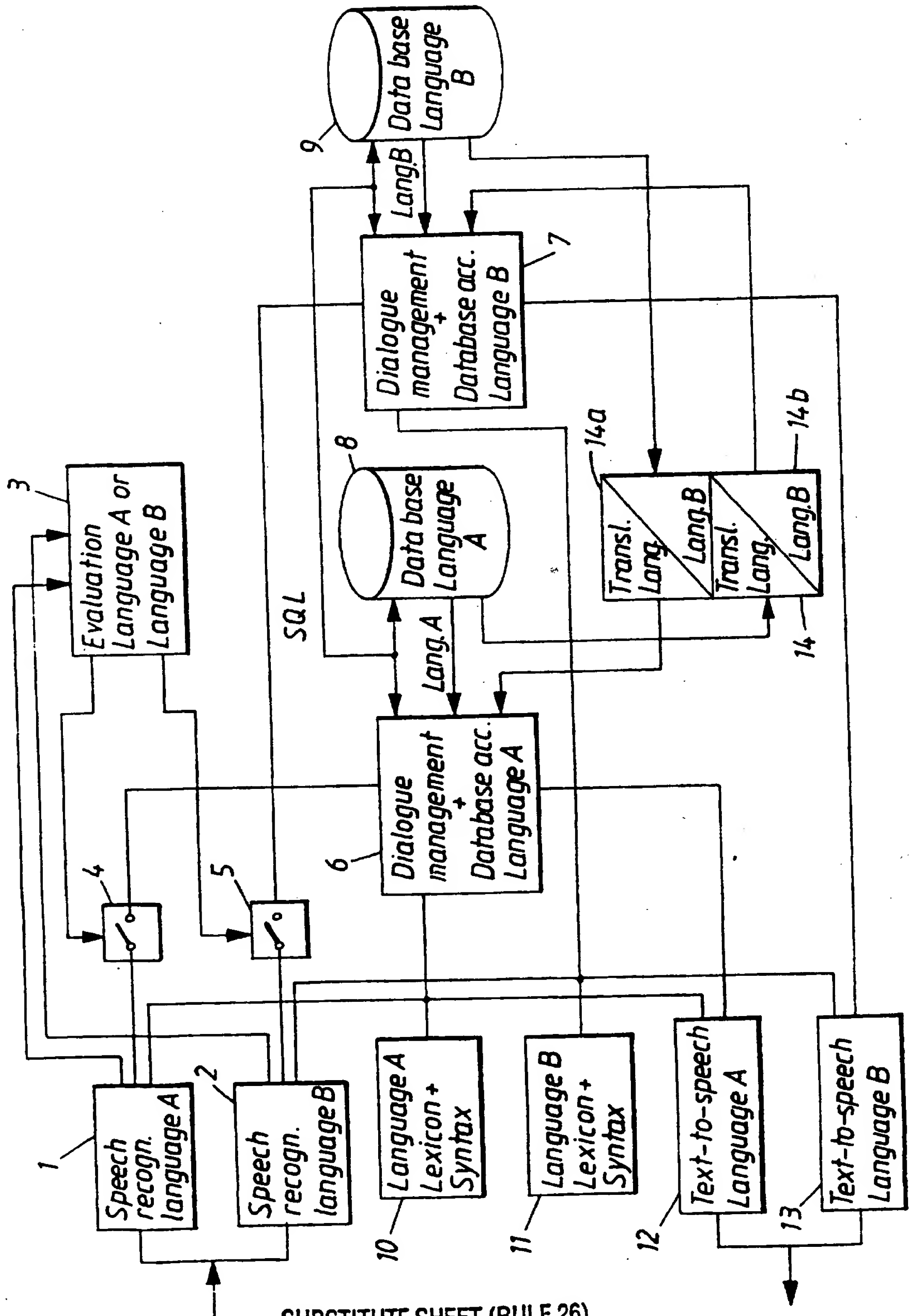
38. A speech-to-speech conversion system as claimed in claim 37, characterised in that the accent information relates to tonal word accent I and accent II.

39. A speech-to-speech conversion system as claimed in any one of claims 36 to 38, characterised in that said sentence accent information is used in the interpretation of the content of the recognised input speech.

40. A speech-to-speech conversion system as claimed in any one of the claims 31 to 39, characterised in that sentence stresses are determined and used in the interpretation of the content of the recognised input speech.

41. A voice responsive communication system including a speech-to-speech conversion system as claimed in any one of the claims 19 to 40.

1 / 1


Speech recogn. language A
Speech recogn. language B
Evaluation Language A or Language B
Dialogue management + Database acc. Language A
Dialogue management + Database acc. Language B
Data base Language A
Data base Language B
Lang. A
Lang.B
SQL
Transl. Lang. Lang.B
Transl. Lang. Lang.B
Language A Lexicon + Syntax
Language B Lexicon + Syntax
Text-to-speech Language A
Text-to-speech Language B
1
2
3
4
5
6
7
8
9
10
11
12
13
14
14a
14b

SUBSTITUTE SHEET (RULE 26)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SE 97/00584

A. CLASSIFICATION OF SUBJECT MATTER

IPC6: G06F 3/16, G10L 5/04

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC6: G06F, G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

SE,DK,FI,NO classes as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GB 2165969 A (BRITISH TELECOMMUNICATIONS PLC), 23 April 1986 (23.04.86)	1-3,8,30
Y		4-7,9-29, 31-41
	--	
Y	EP 0624865 A1 (TELIA AB), 17 November 1994 (17.11.94)	4-7,19-29
	--	
Y	WO 9600962 A2 (TELIA AB), 11 January 1996 (11.01.96)	9-18,31-41
	--	

☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" erlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search 28 August 1997	Date of mailing of the international search report 01 -09- 1997
Name and mailing address of the ISA/ Swedish Patent Office Box 5055, S-102 42 STOCKHOLM Facsimile No. +46 8 666 02 86	Authorized officer Jan Silfverling Telephone No. +46 8 782 25 00

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SE 97/00584

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP 0543329 A2 (KABUSHIKI KAISHA TOSHIBA), 26 May 1993 (26.05.93) --	1-41
A	WO 8903083 A1 (SIEMENS AG), 6 April 1989 (06.04.89) -- --------	1-41

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/SE 97/00584

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
GB 2165969 A	23/04/86	NONE	
EP 0624865 A1	17/11/94	JP 6332494 A	02/12/94
		SE 500277 C	24/05/94
		SE 9301596 A	24/05/94
		US 5546500 A	13/08/96
WO 9600962 A2	11/01/96	EP 0767950 A	16/04/97
		SE 504177 C	02/12/96
		SE 9402284 A	30/12/95
EP 0543329 A2	26/05/93	JP 5216618 A	27/08/93
		US 5357596 A	18/10/94
		US 5577165 A	19/11/96
WO 8903083 A1	06/04/89	DE 3732849 A	20/04/89